Weekly Report

Junhua Lu

June 14, 2015

Gongan

i:Some improvement for regression:(1)normalization of records. (2)temporarily kick out the attribute of birthplace. (3)temporarily we just use the aggregation of one period.

ii:Some statistics:

- For 开房记录, most of the records are in the period of 2011-2013.
- no crime records after 2010. criminal records covers more years, but we had to use R_ZHRQ(抓获日期) as their time of commit a crime. In year2011: 28410 people, year2012: 27088 people, year2013: 26012people.
- Do not exist any record of registration records for temporary resident population(暂住人口). Date of issue(发证日期) is available. In year2011:1437656, year2012:1736460, year2013:2643815.
- using above data, we assume that residents(常住人口) around 8million, temporary resident population 2million, crime rate 2.6
- around three fourths of residents do not have records of occupations.
- There are 32234 crime records, 26453 after duplicate removal. 524886 criminal records, 399259 after duplicate removal. 19542 persons appear both in crime and criminal(after duplicate removal). I did this work to separate good man from bad man since no crime records after 2010.
- In 开房记录, 4256 criminal(2824 after duplicate removal), 390 crime(235 after duplicate removal), 385781 different person has the 开房记录. 49 is the highest records of 开房次数during the three years. As a sample, we extract around 500ciminal for training model, 500 for testing model.

iii:some problems

- Slight difference between different SQL database editor caused problems while processing the data. I used cursor(游标) to assign value to "whether employed or not" and failed many times until I found the PL/SQL software is different in some ways with Microsoft SQL Server. And when I want to pick out the good man in 开房记录, using "Not In", the software crashed

so many times and I was almost getting crazy. They say "not exist" is more efficient than "not in", I will test it again next time I went there.

- Still, some noise data is there remain process. Some data is really beyond imagination, though we may not take it into concern temporarily.

Netease

Fully read the review and will discuss it next Monday.
next week, I will try my best to compute the regression. Besides, I may join in paper reviewing and revising of netease. Also, prepare for another statistical learning slides if free time is enough .